A Dual-Mode Human-Machine Interface for Robotic Control based on Acoustic Sensitivity of the Aural Cavity

Ravi Vaidyanathan & Monique Fargues
Graduate School of Engineering and Science
Naval Postgraduate School,
777 Dyer Road, Monterey, CA, USA, 93943-5139
rvaidyan@nps.edu, mfagrues@nps.edu

Srinivas Kotta & Lalit Gupta
Department. of Electrical Engineering
Southern Illinois University
Carbondale, Illinois, 62901-6603
lgupta@siu.edu

Dong Lin & James West
Division of Rehabilitation Products
Think-A-Move, Ltd
23715 Mercantile Rd., Suite 100, Beachwood, OH, USA 44122
info@think-a-move.com

Abstract **–**We introduce an unobtrusive sensor-based control system for human-machine interface to control robotic and rehabilitative devices. The interface is capable of directing robotic or assist devices in response to tongue movement and/or speech without insertion of any device in the vicinity of the oral cavity. The interface is centered on the unique properties of the human ear as an acoustic output device. Our work has shown that various movements within the oral cavity create unique, traceable pressure changes in the human ear, which can be measured with a simple sensor (such as a microphone) and analysed to produce commands signals, which can in turn be used to control robotic devices.

In this work, we present: 1) an analysis of the sensitivity of human ear canals as acoustic output device, 2) the design of a new sensor for monitoring airflow in the aural canal, 3) pattern recognition procedures for recognition of both speech and tongue movement by monitoring aural flow across several human test subjects, and 4) a conceptual design and simulation of the machine interface system.

Index Terms **–** Ear pressure signals, human-machine interfaces, signal classification, signal detection, signal estimation

I. Introduction

A well-recognized need exists for tools enabling the physically impaired to be more independent and productive. In general, devices developed to assist the disabled involve detecting an input signal produced by the user and converting that signal into an electronic command signal, which in turn causes a desired event to occur. However, the majority of mechanisms designed for human use require the user to generate input signals through bodily movements, most often with their hands, arms, legs, or feet. Such devices clearly exclude individuals with impairments that cause painful or limited control of their appendages.

The goal of our ongoing research is to develop a human-robotic interface which can overcome the deficits of these systems for seamless operation of mobile platforms, for assist device control in particular. In this work, we introduce a method for detecting both tongue movement and speech, and generating a control instruction corresponding to that action that can be applied to any tele-operated or semi-autonomous robot. We have previously reported on the development of a non-intrusive tongue-movement based machine interface without the need for insertion of any device within the oral cavity [3, 4]. This interface consists of tracking tongue movement by monitoring changes in airflow that occur in the ear canal. Tongue movements within the human oral cavity create unique, subtle pressure signals in the ear that can be processed to produce commands signals in response to that movement. Once recognized, said movements can in turn be used in for robotic tele-operation.

In this work, we expand on our past research to present: 1) an analysis of the sensitivity of human ear canals as acoustic output device, 2) the design of a new sensor for monitoring airflow in the aural canal, 3) pattern recognition procedures for recognition of both speech and tongue movement by monitoring aural flow across several human test subjects, and 4) a conceptual design and simulation of the machine interface system.

II. Machine Interface System

Our system is designed to provide a smooth and accurate method for a human to communicate, command, and control devices through tongue movement and speech commands. The interface consists of monitoring air pressure within the human ear, and subsequently providing corresponding control instruction. The system makes use of changes in air pressure or sound waves (vibrations) in the ear to characterize measured parameters. Research has shown that initiating actions, in particular movements of the tongue [3, 4] and speech, produce detectable pressure waves with strength corresponding to the direction, speed and/or intensity of the action*..

[*] Laboratory evidence [5] suggests that thought and intention may be detected and recognized as well. This potential will be addressed in future work.

Our research has shown that initiating actions in the oral cavity generate air pressure changes in the range of 10 Hz to 4 kHz. Specifically, tongue movements normally are traced between 20 Hz and 90 Hz while speech normally occurs between 250 Hz up to 4 kHz. Other initiating actions, such as singing and biological processes, generate air pressure changes in the range of about 20 Hz to about 20 KHz.

Figure 1 illustrates a sensor inserted partially into the ear of a person (i.e. within the cavity defined by the pinna, if not deeper within the ear such as within the concha, at the opening of the ear canal). The sensor includes a housing and internal microphone. The illustrated housing is made from a material such as plastic and is wider than the opening of the ear canal, so as to engage the pinna or, alternatively, to cover the whole pinna (e.g., similar to a hearing aid or cell phone communication device).

The interior portion of the housing has a recess in which the microphone is placed. By inserting the microphone in the ear and/or ear canal, the microphone is shielded from environmental noise. The microphone is capable of detecting various forms of initiating actions, including, for example, physical movements of the user such as touching the tongue lightly in different parts of the mouth, touching the tongue to certain parts of the mouth, or any combination thereof.

III. Modeling of air flow within the Human ear canal

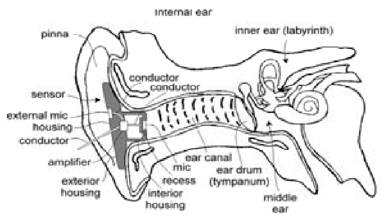

Figure 1: Structure of Ear and Device Insertion

A. *Ear canal pressure change due to its volume variation*

The ear canal is modeled as a 2 cm3 volume. When tongue or cheek moves, forces will be created around the walls of ear canal, which in turn changes the volume of the ear. The whole process is approximated as adiabatic. Therefore, according to thermodynamics theory, we have

$$PV^{\gamma} = C \quad (1)$$

where P is the air pressure in the ear canal, being the summation of atmospheric pressure (P0) and induced acoustic pressure (p); V is the volume of the ear canal, being the summation of the static volume and the variation due to tongue movements; γ = 1.4 being the specific heat ratio of the air; and C is a constant. The relationship between the induced acoustic pressure and the variation of ear canal volume can be obtained from Eq. (1).

$$p = -\gamma P\frac{\delta V}{V} \approx -\gamma P_0\frac{\delta V}{V_0}; \quad (2)$$

Where p is the induced acoustic pressure in the canal, δV is the volume variation of the canal and V_0 is the static ear canal volume.

Figure 2A demonstrates the effect of the variation of ear canal volume to the acoustic pressure inside the canal. Notice that a relative change of canal volume of one millionth introduces an acoustic pressure of 77 dB ref 20 µPa. Thus, with a volume variation of only 0.002mm3 in our model, a significant acoustic pressure is created inside the ear canal, justifying our premise of using it as a sensitive acoustic output device.

B. *Ear canal acoustic pressure due to volume speed of airflow*

Section III-A enumerated the pressure variation of the ear canal statically. In actuality, the volume change within the ear canal varies dynamically according to the volume speed of airflow. In the following analysis, the volume speed of airflow is treated as a harmonic signal with amplitude of v/2 and an angular frequency of ω.

At low frequency the equivalent circuit for the 2cm3 coupler is treated as an acoustic capacitance. Therefore, the induced acoustic pressure inside the ear canal due to volume speed of airflow is:

$$p = \frac{v\gamma P_0}{\omega V}; \quad (3)$$

Where v is the volume velocity, γ = 1.4 is the ratio of specific heat of the air, V = 2cm3 is the volume of a standard coupler, and ω is the angular frequency of the volume speed when being treated as harmonic vibration.

Figure 2B shows the effect of volume speed on ear canal acoustic pressure at frequencies of 10Hz, 100Hz, and 1 kHz. For v being 1mm3/s and f being 10Hz, the acoustic pressure created is p =1.13 pa, which corresponds to 95.0 dB ref 20µPa. Both models demonstrate the sensitivity of the ear canal as an acoustic output device.

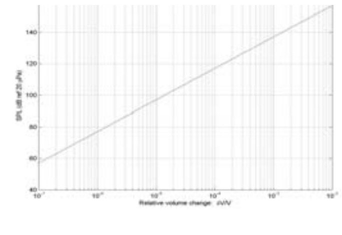

Figure 2A: Pressure vs. Change of Ear Canal Volume

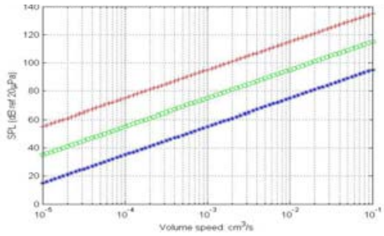

Figure 2B: Pressure vs. volume speed (red @ 10Hz, green@100Hz, blue@1kHz)

IV. Sensor (Earpiece) Design for Aural Flow Monitoring

Our research team has designed and, through iterative prototypes, significantly improved the performance of the earpiece-sensor housing to detects pressure fluctuations in the ear canal. The first generation earpiece, pressure sensor, and circuitry was housed in a custom molded earplug housing. This first generation housing was similar to a plug used in hearing aids. Figure 3 shows a picture of our custom-designed microphone-earpiece housing next to an actual mold taken of the left ear of the test subject, and a photograph of a test subject comfortably wearing the ear-piece microphone housing. The ear piece (shell) was made from the mold on the left. Although the system performed quite well, the need for a custom earpiece limited the utility of the device and the size of the device raised questions with respect to system robustness.

Our research team has developed a second generation physical housing suitable for use with a wider range of subjects with little or no customization. The result is the earpiece shown in Figure 4. The new earpiece system is separated into two components. The portion of the device that is actually inserted in the ear to pick up pressure fluctuations is a soft foam shell with a tube that connects the ear canal to the sensor and electronics housing. Studies conducted for sensor placement (based on acoustic air flow models developed enumerated previously) dictated the shape and depth of insertion of the microphone-ear piece housing. The microphone resides on the interior portion of the housing within the ear canal at a depth of 2.5 mm to 12.5 mm measured from the opening of the ear canal. The sensor and electronics housing are formed into a small molded shell, which is then fitted over the back of the ear. The resulting second-generation system has been demonstrated to provide comparable performance and comfort to the first generation system and is more easily adaptable to a wide range of users. Furthermore, due to the compliant soft foam insertion, the new earpiece enjoys even greater benefits with respect to shielding pressure signals from environmental noise.

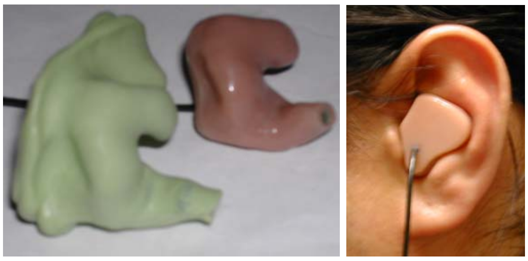

Figure 3: First Generation Sensor for Signal Capture

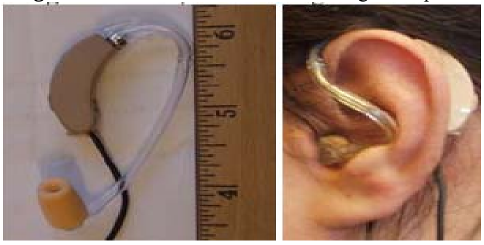

Figure 4: Second Generation Sensor for Signal Capture

V. Measure of Aural Flow Resulting from Initiating Actions

A. Tongue Movement

Based upon extensive feedback from test subjects, we have defined four basic tongue movements for robotic interface, which nearly all patients should be capable of generating. These are: touching the tongue to the top/front center of the roof of the mouth, and "flicking" it gently forward ("forward" movement), touching the tongue to the bottom/front center of the mouth, the front/right side of the mouth, or the front/left side of the mouth and "flicking" it gently up from any of these positions ("backward", "right", and "left" movements). "Backwards", "right", and "left" tongue movements are illustrated graphically in Figure 5. We therefore refer to this set of 4 movements as the "standard" interface.

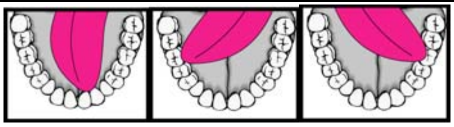

Figure 5: Graphic of Three Tongue Initiating Movements

Figure 6 shows a sample of raw data gathered from a microphone embedded in the housing described earlier, and inserted in the ear of a subject as shown in Figure 4. The subject was asked to make a "right" movement described earlier. As can be seen from the figure, a very clear change in microphone output is seen, which corresponded directly to the movement of the tongue.

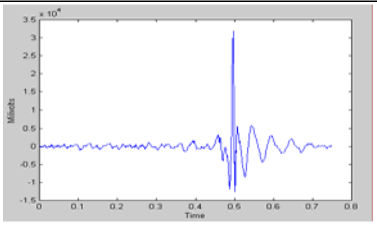

Figure 6: Tongue Movement Data

B. Speech

Figure 7 shows speech data collected by the second generation sensor in a highly noisy environment. Figure 7A shows data collected with the sensor located in the ear, while Figure 7B shows data collected with the microphone located in front of the mouth (with the same word and background noise) for comparison to traditional speech recognition. These two plots clearly illustrate the noise shielding capability of the device when inserted in the ear, which points the superiority of the device in highly noisy environments compared to other speech recognition systems. While other research groups [1, 2] have investigated speech capture in the aural cavity, this only work we are aware of that has made use of a non-customized sensor.

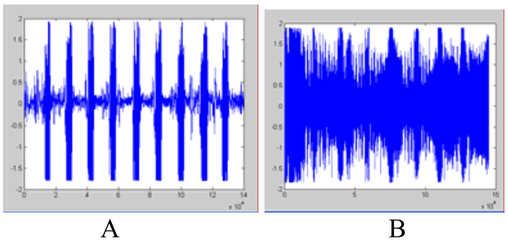

Figure 7: Aural Speech data, 9 trials of the word "one", high noise environment; A) Sensor located in the ear; B) Sensor located in front of the mouth.

VI. Signal Recognition Procedure

A. *Tongue Movement Recognition*

At this time, we have developed a strategy to accurately detect and classify, in real time, changes in the air flow pressure that occur in the ear-canal caused by tongue movements, or Tongue-Movement Ear-Pressure (TMEP) signals. At this time, we have developed a unique Decision Fusion Classification Architecture capable of classifying the previously described 4 "standard" movements with over 96% accuracy. Details of this scheme for tongue movement are enumerated in [7]. Table 1 shows recognition accuracies for 4 test subjects of varying ages using the 4 movements in the "standard" interface. The numbers shown illustrate the percentage of the time that TMEP signals resulting from each of the 4 movements were successfully distinguished by the Decision Fusion Classification Architecture from the other 3.

Subject	Recognition Accuracy
S1 (male, 43)	98.26
S2 (male, 33)	97.25
S3 (male, 21)	98.46
S4 (female, 54)	99.68

B. *Speech Recognition*

For speech recognition, we have developed a Hidden-Markov Model (HMM) algorithm. Hidden-Markov Model (HMM) classifiers have been used extensively to model the temporal structure and variability of signals, especially in speech applications over the last few decades. Our specific application considers a small dictionary of isolated words and a basic discrete HMM classifier set-up was investigated for that task. Details of its implementation will be enumerated in future publications [REF!].

We selected a discrete HMM structure, as it has been shown to lead to satisfactory results in isolated word recognition and is simpler to implement than a continuous HMM structure. Therefore, we applied a vector quantization scheme to map the sequence of continuous valued feature coefficient vectors into a sequence with a given number of discrete vectors, called the codebook so as to generate a finite set of permissible feature vectors using the K-means algorithm.

Given that the purpose of our interface is human-machine interface, seven monosyllabic words were selected in our initial study, based on commands a user might give to a robot or assist device (such as a power wheelchair). The words chosen were: {up, down, left, right, move, pan, kill}; three adult speakers were tested in this first study: 2 females and 1 male. Initial data was collected in a quiet office environment and data sampled at 8KHz. Resulting average classification performances are shown in the table below in Table 2, where the classifier was run three times by varying the order in which the three subjects were presented to the classifier. The left column indicates the voice action input by the subject, while the top row indexes the resulting classification assigned to the input by our pattern recognition algorithms. For example, when the subject gave (spoke) the robot a 'kill' command, the system correctly recognized the command 98.89% of the time, mistook it for an 'up' command 1.11% of the time, and a 'right' command 1.11% of the time.
These results are encouraging as they indicate an average classification performance equal to 95.87% for the seven words considered. Furthermore, all commands from all users may be recognized universally by the HMM classifier, thus providing firm evidence that capture of speech in the ear can be accomplished with calibration comparable to existing speech recognition devices.

Performance (%)	up	down	left	right	move	pan	kill
up	94.45	1.11	1.11	1.11	1.11	1.11	1.11
down	0	90	0	6.7	0	3.33	0
left	1.11	0	93.33	5.56	0	0	0
right	0	1.11	0	97.78	0	0	1.11
move	0	0	0	0	100	0	0
pan	0	1.11	0	2.22	0	96.67	0
kill	1.11	0	0	1.11	0	0	98.89

Table 2: Average Classification Accuracy: 95.87%

VII. Conceptual Design of Machine Interface

A. *Device control through speech interface*

As stated earlier, seven words were selected in our initial study: {up, down, left, right, move, pan, kill}. These words can be implemented as a set of commands a patient may give to an assist devices (e.g. a power wheelchair), or a mobile robot with a camera. While more complex schemes were possible, at this time we have implemented a first generation conceptual design of the interface system for control of an assistive robot or power wheelchair. We propose a straightforward system designed around four words for motion control. These are centered around the words 'up', 'down', 'left', and 'right'. These four words can be coupled to create an intuitive interface such that a 'right' command corresponds to a right movement, with 'left' movements following naturally. Finally, in the proposed interface, a 'kill' command executes an all stop command. All enumerated commands would be very straightforward to implement in a standard communication setup.

Open Environment Simulation

A version of this control interface was implemented in simulation to prove the current speech recognition accuracies are sufficient for robotic control. In this interface, an “up” command movement was assigned to move the robot forward. A “down” movement stopped the robot if given when the robot was moving forward. If passed to a stationary robot, this same command would move the robot in reverse. Intuitively, a “right” or “left” movement altered the robot’s heading by 5° in either direction respectively. Speech recognition errors were included based on the accuracies presented in Table 2. A 0.3-0.4 second interval t delay was assumed between movement commands.

Figure 9 shows the results of a simple simulation where the interface was implemented to direct a robot (with a forward velocity of 1 m/s) † to reach a series of (20) waypoints in a planar work space. The “+” symbols represent the waypoints with the path of the robot shown. The waypoints were spaced arbitrarily across a 200m amplitude sinusoidal path with a period of 80m. In each case, a “virtual” operator was provided the planar position of each successive waypoint and the robot’s position. In order to assess the impact of any erroneous operator commands, this particular simulation was repeated 1000 times. In every case, the robot successfully reached all waypoints without fail. While some commands were mistaken by the interface over each run, the high rate of accuracy and speed at which commands may be given allowed immediate correction for any mistaken commands.

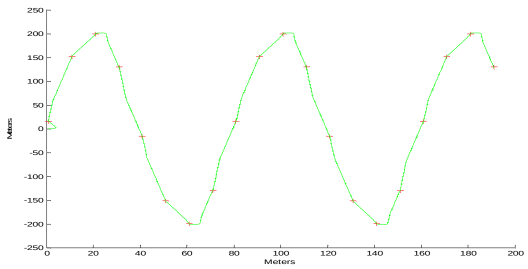

Figure 9: Robot Waypoint Navigation

B. Device control through tongue movements

The four movements in the “standard” tongue-based interface were considered in a conceptual interface for device control. A top touch could move the robot forward, and left and right touches could increment portions of turns. A bottom touch could be a complete stop. The system would thus be intuitive; a movement of the tongue to the left would cause the robot to turn left, etc. The speed, however, would be a constant unless an alternative system design was used where each additional (top) touch would increment speed slightly, and a bottom touch would decrement speed slightly. In this case two immediate touches to the bottom could be a full stop. Another possibility would be for each top touch to put a speed "pulse", or square wave into the system, which will gradually decrease over a period of time. Two bottom touches may serve as an "all stop", while other sets of touches could activate menus for non-movement based activity. Beyond forward/reverse and left/right motions, additional commands may be necessary to control the robot. In order to correlate robot actions to additional movements, a time signature between movements is proposed. In this system, any movement repeated within a fixed time period (Δt) of the same preceding movement may be considered to be a separate movement. For example, a top movement followed by another top movement within a period t<Δt will correlate to a different action than a top movement followed by a pause>Δt, and another top movement. Based on present data and user feedback, setting Δt = ~0.2 seconds is comfortable to most users. Note that Δt may be smaller if more rapid control is required; it is relatively easy for a user to repeat the same movement within a very short time frame. TMEP recognition accuracies for each user were used in a similar manner to the previous simulations to provide robot performance specifications.

Functionally, such an approach is quite similar to a sip-and-puff tube controller commonly used by quadriplegics to control wheel chairs, which we studied for inspiration in our system design. Patients sip or puff to turn in discrete increments, and sip or puff to increase or decrease speed. The advantage of our system over this traditional system, allows a greater potential range of inputs, the enacting movements are easier to perform, the time scale is quicker, and there is no need to place any device within the oral cavity. .

Constrained Environment Simulation

A “crowded” environment was used as a simulation testbed for a feasibility proof of concept for robot control. Beyond accurately maneuvering in open environments, robots, particularly those envisioned as patient assist devices must often perform fine control in crowded environments. While collisions with obstacles occasionally occur with virtually all existing interfaces, it is critical that they be kept to an absolute minimum. Thus, we conducted a series of simulations designed at testing the ability of our interface for fine maneuvering in cramped environments.

A virtual “obstacle course” was created consisting of an environment with several obstacles of various sizes placed at arbitrary locations. A virtual operator was then tasked to maneuver the robot through the room using the standard control interface described earlier. Signal recognition accuracies for each movement were used for subjects S1, S2, S3, and S4 (detailed in [7]) to provide a realistic appraisal of the robot’s performance. Figure 10 shows the results of one such simulation. In the simulation shown, a robot under the control of a test subject was placed in an environment comprised of a variety of obstacles forming a narrow canyon only slightly wider than the vehicle itself, similar to the interior of a home.

† The robot performance parameters selected for this simulation were based on the “Whegs II” robot constructed at Case Western Reserve University [8]

Of importance to note, is that while some tongue movement commands were mistaken by the system (approximately 20 commands were identified incorrectly in the simulation shown for test subject S1), and despite the corridors the robot maneuvered through, no collisions between the robot and obstacles were recorded in the simulation. The high rate of recognition accuracy and speed at which commands may be given allow for immediate correction for any mistaken commands, thus all potential collisions may be avoided The same simulation was repeated 1000 times with data from all test subjects. Collisions were recorded less than once per 1000 runs for every test subject. The nature of the interface system coupled with the extremely remote possibility of any repeated error allows for virtually error-free operation, even in restrictive environments. Furthermore, in the very rare event of a collision, resuming the original path is a very easy task.

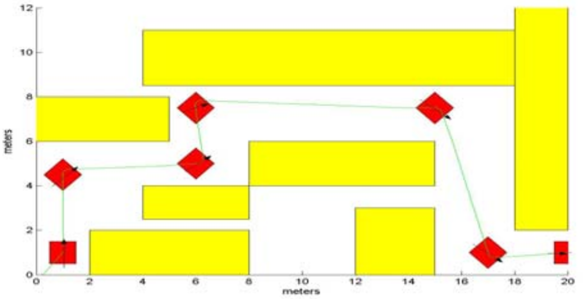

Figure 10: Robot Constrained Environment Navigation

As a final test aimed to understand the control system's ability to correct erroneous commands, a series of simulations were run with induced errors in the pattern recognition strategy. For example, in one case, errors were induced to reduce the pattern recognition of a "left" movement to 80% accuracy with the principle recognition error being a "right" movement. When this system was implemented in the same simulation shown in Figure 10, approximately 10% of the trials resulted in at least one collision. Thus, even with radically reduced recognition accuracies, the consequences of misrecognized commands still rarely occur in a collision.

VIII. Conclusions

The goal of this paper was to expand upon a new concept for hands-free communication and control in human machine interface, particularly in the assistive device arena. To our knowledge, our research team is the only group that has investigated the aural cavity as a monitoring venue for machine interface, and has proposed the only system whereby tongue movement and speech may be tracked without insertion of any device in the oral cavity.

In this work we report the development of a bi-modal human machine interface capable of tracking both speech and tongue movements in a single unobtrusive device. Each method has complementary strengths which could be synergized in a comprehensive system. We have observed tongue movement to be faster, quieter, and (in most cases) more intuitive to the user for direct device motion control when compared to speech. Aural speech capture, however, does require less calibration and training on the part of the user, and may have the potential for a wider range of inputs.

Future work involves synergizing both the speech and tongue movement modes of interface to develop a cohesive, robust human/robot interface that will allow one to control and task robotic platforms in any situation without causing additional weight, and without the addition of any bulky or encumbering equipment. In the longer term, two distinct modes of operation with the device are envisioned whereby several devices (e.g. a power wheelchair, household appliances, stationary mechanical assist devices, etc.) may all be directed due to the infinite possibilities for control .input. We believe this work will lay the foundation for a new generation of hands-free human machine interface systems for all manner of rehabilitative and robotic application.

IX. References

1. Westerlund, N., Dahl, M., and Claesson, I., "In-ear Microphone Equalization Exploiting an Active Noise Control," Proceedings of Internoise 2001, August 2001.
2. Westerlund, N.,Dahl , M., and Claesson, I., "Speech Recognition in Severely Disturbed Environments Combining Ear-mic and Active Noise Control," Internoise 2002, August 2002.
3. R. Vaidyanathan, H. Kook, L. Gupta, & J. West, "Parametric and non-parametric signal analysis for mapping air flow in the ear canal to tongue movements: A new strategy for hands-free Human-machine interface," IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004, Montreal, Quebec, Canada.
4. R. Vaidyanathan, T. Huynh, T. Allen, R.D. Quinn, M. Tabib-Azar, J. Zarycki, J. Levin, B. Chung, & L. Gupta, "Human-Machine Interface for Tele-Robotic Operation: Mapping of Tongue Movements Based on Aural Flow Monitoring," IEEE International Conference on Intelligent Robots and Systems (IROS), October, 2004.
5. G. G. Nemisrovski and G. L. Troussov, "System and Method for detecting a thought and generating a control instruction in response thereto," Patent US6024700; Issued 15 Feb, 2000.
6. American National Standards Institute (ANSI) S3.7 and IEC 711 standards
7. R. Vaidyanathan, S. Kotta, L. Gupta, & J. West, "A Decision Fussion Classification Architecture for Robotic Interface: Mapping of Tongue Movements Based on Aural Flow Monitoring", Unpublished, Please contact corresponding author (rxv@case.edu to obtain)
8. Quinn, R.D., Kingsley, D.A., Offi, J.T. and Ritzmann, R.E., (2002), "Improved Mobility through Abstracted Biological Principles", IEEE Int. Conf. On Intelligent Robots and Systems (IROS'02), Lausanne, Switzerland